DOCUMENT RESUME

ED 286 907 TM 870 539

AUTHOR Gilmer, Jerry S.
TITLE The Effects of Test Disclosure on Linear Equating Relationships under the Common Item Nonequivalent Groups Design.
PUB DATE Apr 87
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adults; Answer Keys; Computer Simulation; *Cutting Scores; Difficulty Level; *Disclosure; *Equated Scores; Licensing Examinations (Professions); State Legislation; Testing Problems; *Test Items; Test Validity
IDENTIFIERS Anchor Tests; Fairness; *Test Disclosure; *Test Security; Truth in Testing Legislation

ABSTRACT

The proponents of test disclosure argue that disclosure is a matter of fairness; the opponents argue that fairness is enhanced by score equating which is dependent on test security. This research simulated disclosure on a professional licensing examination by placing response keys to selected items in some examinees' records, and comparing their scores on a later form of the same test to scores of examinees who did not receive disclosure. Ten groups varied in number of items disclosed, number of examinees receiving disclosure, ability level of the subgroup receiving disclosure, and whether the items disclosed were anchor test or nonanchor test items. Results depended on the degree of exposure to test items; the greatest score differences were found in the group in which the greatest number of people received disclosure on the greatest number of test items. The effect of item disclosure on passing scores and passing rates was also examined. Results depended on whether the disclosed items were anchor or nonanchor items and whether the benefit of disclosure was direct or indirect. It was a concluded that there is currently no perfect equating method for test disclosure and that test fairness is more complicated than originally thought. Efforts to have fair disclosure in occupational licensing tests must be weighed against the need to protect the public. (JGL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED286907

The Effects of Test Disclosure on Linear Equating Relationships Under the Common Item Nonequivalent Groups Design

Jerry S. Gilmer

Office of Consultation and Research in Medical Education

The University of Iowa

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Gilmer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper Presented at the Annual Meeting of the
American Educational Research Association
Washington, DC, 1987

The author is grateful to The American College Testing Program for support received on this project, with special thanks to Mr. Fong-Ching Cheng at ACT for computer assistance.

TM870539

BEST COPY AVAILABLE

Abstract

The proponents of test disclosure argue that disclosure is a matter of simple fairness; the opponents argue that fairness is enhanced by score equating which is dependent on the security of the test. This research involved simulating disclosure by placing correct answers of "disclosed" items into response vectors of selected examinees. The degree of exposure the disclosed items received in the population was manipulated by varying the number of items disclosed and the number of examinee records receiving the correct answers. Other factors considered among the ten experimental conditions included the characteristics of the disclosed items (difficulty of disclosed items and whether they were anchor test or nonanchor test items) and the ability level of the subgroup receiving the disclosed items. Results suggest that effects of disclosure are dependent on the nature of the released items and the conclusions discuss the issue of fairness to examinees and to the general public.

The Effects of Test Disclosure on Linear Equating Relationships Under the Common Item Nonequivalent Groups Design

Introduction

The debate and the controversy surrounding truth-in-testing legislation have centered on the issue of test disclosure. A requirement of test disclosure implies that examinees be given an opportunity to see the test items after the test administration. The essence of the proponents' arguments for test disclosure is consistent with values that are part of the foundation of American democracy; "The argument is that this is a matter of truth, fairness, and open governance" (Brown, 1980, p. 53). The basic argument of the opponents of test disclosure is that test security is essential to test validity, and a threat to the validity of a test is a threat to the fairness of the test. When evaluated more closely it becomes apparent the two sides are not arguing the same issues on the same level. Brown writes:

> When the major groups of arguments are examined it becomes obvious that the antagonists are seldom drawing upon the same facts, looking at the same aspects of the educational enterprise or subscribing to the same beliefs about the nature and function of education. *They are more often arguing at each other than with each other.* (Brown, 1980, p. x, emphasis added.)

The prospects of this controversy evolving to a level where the sides are arguing on common ground are limited by the nature of the arguments and the extreme scarcity of empirical research. It is difficult to be directly opposed to either open governance or valid testing. The issues and the values embedding these two points certainly do not readily lend themselves to empirical investigation. There is, therefore, strong motivation and, hopefully, potential in attempting to move the controversy away from issues that are difficult to evaluate and into the scientific arena. The primary purpose of this paper is to attempt to take a small step in moving the controversy into a more scientific context. The paper will present a brief background of truth-in-testing legislation and the major points from each side of the test disclosure issue. The need for more scientific inquiry will be

supported with reference to more extensive and fairly objective summaries of the issues. The results of several conditions of simulated test disclosure will then be presented and discussed. The relationship between test disclosure and fairness will be examined. Finally, some of the tentative conclusions of the research will be briefly discussed.

Background

In the late 1970s and early 1980s the concept of "truth-in-testing" received intense and widespread attention when several states and even the U.S. Congress were considering legislation which would seriously impact the development, use and interpretation of tests in this country. Between 1977 and 1983 approximately 90 testing bills were introduced in 28 states; five bills were introduced in the U.S. House of Representatives (Greer, 1984a). (It is likely that the number of bills introduced includes the reintroduction of identical or similar bills in subsequent legislative sessions.) Although only two states--California and New York--have enacted testing laws, the significance of these two states to several national testing programs has led to de facto truth-in-testing on a national basis (Greer, 1984b, p. 329). The testing laws in California and New York and most of the other bills introduced apply only to tests used for admission to postsecondary education and, therefore, the issues involved are of obvious concern to all those who work with admissions tests. The issues are also important to sponsors of occupational licensing tests because of the increased pressure being applied to sponsors and licensing boards by their various constituencies. The laws require test agencies to disclose certain technical information such as research and validity reports and rules for converting raw scores in addition to requiring disclosure of the test items. The debate, however, has focused on test item disclosure as the primary issue.

The main argument of the proponents of test disclosure is that examinees are entitled to know the basis upon which they are being judged, as a matter of simple fairness (Strenio,

1979; Brown, 1980). The following arguments, contained in Brown's (1980) summary of the fairness issue, represent the essence of the position:

> . . . in a free society people should know exactly how they have been evaluated and judged. Nothing less should be tolerated. (p. 27)

> . . . open testing is more fair to our society, which being democratic, should prefer governance in the open. (p. 54)

In summarizing the fairness argument, Strenio (1979) includes the following:

> If test takers are expected to abide by the results with equanimity, they should be able to examine and evaluate the test contents. (p. 9)

The opponents of disclosure argue that fairness is already a major concern in the development and use of tests. Brown (1980) writes:

> Security is necessary, they argue, in order to ensure that some students do not have unfair advantage over others; to equate tests over time; . . . (p. 29)

Strenio (1979) presents the following argument:

> In essence secure testers claim that releasing test questions would greatly endanger both current methods of validating standardized tests and also the statistical equating of test results over time. (p. 17)

The purpose of test equating is to ensure that a specific score received on one form of a test has exactly the same interpretation as the same score received on another form of the test. For example, two examinees who are at the same level of achievement, relative to the content of the test, should receive the same score even if they are administered different forms of the test. If the test forms are not equated, the examinees' scores may not be the same due to unintended differences in the difficulty levels of the two forms; one form may be easier than the other form. Under equating, adjustments will be made in the raw scores to remove the effects of different difficulty levels and the two examinees will receive the same score.

Often in test score equating an "anchor test" is used to determine the equating relationship (i.e., to determine the appropriate adjustments). The anchor test is basically a

short version of the test forms to be equated and is administered with those forms. One form may be administered several months or years after the other form. The reasoning underlying the argument of the opponents of disclosure is that if the items on the first form are disclosed there is likelihood that the examinees who will take the second form will see the items in the anchor test and receive scores that are spuriously high as a direct result of the disclosure of the first form.

The logic behind the opponents' argument is difficult to refute. But the lack of scientific studies examining the issues is cited in the summaries of the controversy. In drawing his conclusions Strenio points out that "on many key issues there is little evidence available on which to base a judgment" (Strenio, 1979, p. 54). Brown writes that the relationship between test security and test validity remains an unexplored area and that "[t]he debate could profit from a clear explanation of this point on a test by test basis" (Brown, 1980, p. 54). Brown also states that "[t]he proposition that disclosure necessarily leads to lower quality tests . . . needs more comprehensive testing before it can be said to be proven" (Brown, 1980, p. 55). In 1978 the National Academy of Sciences convened the Committee on Ability Testing to evaluate the role of testing in American life. The following statement represents one of the Committee's views:

> Speculation about the short- and long-range consequences of test disclosure should be replaced by objective research on the effects of various disclosure plans. (Cited in California Department of Consumer Affairs, Central Testing Unit, 1983, p. V-4)

In the research review for this paper it became evident that little has been reported dealing with the effects of test disclosure. One study examined the effects of disclosure on examinee performance on the Test of English as a Foreign Language (TOEFL) (Hale, Angelis, & Thibodeau, 1981). The researchers concluded that the disclosure of test items resulted in an increase in scores on another TOEFL which contained the disclosed items. Based on the above quoted calls for research examining various disclosure plans on a test by test basis, it seems clear that additional studies are required. The research that is

discussed below is an attempt to respond to this void. The purpose of the research was to examine the effects of various disclosure conditions on examinee scores for a nationally administered licensing exam in which the scores are equated.

Methods

The basic strategy of this research was to simulate the disclosure of items from one test form, called form A in this paper, to examinees who take a subsequent test form, form B, and then examine the resultant equated scores of the form B examinees. Forms A and B share an anchor test. The number right scores (raw scores) obtained from form B are transformed (equated) to the scale for form A. To examine the effects of disclosure, the equated scores obtained on form B after disclosure were compared to a baseline data set which is simply a set of equated scores for form B with no disclosure.

The test used is a nationally administered professional licensing test that is taken by over 50,000 examinees per year. Form A and Form B were administered three years apart. The test contains 200 multiple choice items of which 30 constitute the anchor test. The anchor test in this testing program is an internal anchor which simply means that responses on the anchor test are part of examinees' raw scores.

Item disclosure was simulated by placing the keyed response to an item selected to be "disclosed" in several examinees' records. If an examinee's original response was the keyed response, no change was made. This method realistically simulates the exposure of specific items to specific examinees; if an examinee knew the correct answer to an item before he or she received disclosed information, the examinee's record would already contain the correct answer and the simulated disclosure (i.e., placing correct answers in a record) would have no effect on that item for that examinee.

A random sample of 5000 examinee records was selected from the national group of examinees who were administered form B. The baseline equating results were obtained by equating the raw scores for these 5000 examinees to test form A without modifying any

examinee records. All equating results were determined through a procedure known as the Tucker method for common item linear equating with nonequivalent populations (Angoff, 1971; Braun & Holland, 1982; Kolen, 1985; Brennan & Kolen, 1986).

From a practical perspective, the specific effects of item disclosure are the results of a complex combination of factors including such variables as the number and characteristics of the disclosed items, and the extent of exposure the items receive in the examinee population. Because of the complex nature of these variables it is understood that it would be very difficult, if not impossible, to simulate disclosure conditions that strictly reflect the true nature and "behavior" of actual disclosed materials. The disclosure conditions created for this research are intended to represent the disclosure of items in test form A to the examinees who were administered form A. It is assumed that through what may best be termed "osmosis" some of the information received by the form A examinees is diffused to some of the examinees who will take form B. The degree of effective exposure which results from this diffusion process may be small--a small number of disclosed items received by a small group of form B examinees; or much more extensive--a large number of released items received by a large group of form B examinees.

The conditions. In addition to the baseline condition which represents no item disclosure, 10 conditions were created representing various combinations of the number of disclosed items and the size of the group receiving the items. These 10 conditions are described below.

1. A1010.1. All conditions with the prefix "A" indicate the disclosure of a random subset of *anchor test* items. The next two digits indicate the percentage of items disclosed. A "10" in this case means that 10% of the anchor test items were disclosed. The next two digits indicate the percentage of baseline examinees (n=5000) who received the direct benefit of the disclosed items. The digit on the right side of the decimal point simply designates a variant of the basic condition. In this condition three of the anchor test items were randomly selected and directly benefited 500 of the 5000 form B examinees. This

condition is intended to represent a relatively small degree of exposure of disclosed anchor test items.

2. A1010.2. The difference between A1010.1 and A1010.2 is that new samples of disclosed items and examinees were randomly chosen. The sampling was with replacement meaning there could be overlap in the items or the examinees chosen for conditions 1 and 2. This condition also represents a relatively small degree of item exposure and was specified in order to help determine if major differences in results would occur simply by randomly selecting different items and different examinees.

3. A1010.3. This condition is similar to the first two conditions but contains a major variation. Ten different groups of 50 examinees received the benefit of 10 subsets of three anchor test items. One group of 50 examinees received three of the anchor test items; another group received three other anchor test items, and so on. In all, 500 examinees each received the benefit of some set of three anchor test items. This condition was specified to represent a disclosure situation which, from a certain perspective, could be more realistic than conditions 1 and 2; not all examinees who receive disclosed information will necessarily receive (or remember) the same information.

4. A1010.L1. The "L" in the title of this condition indicates "low" ability examinees. Five hundred examinees were randomly selected from a subset of the total group which consisted of all examinees whose raw score was below 128, the raw score mean of the total group. A disclosure situation of this nature--primarily low ability form B examinees receiving the benefit of disclosed anchor test items--may occur if only low ability form A examinees are given the opportunity to examine the items in form A. It seems logical to expect greater results under this condition than under the previous three conditions because, intuitively, the lower scoring examinees have more to gain. Under a situation of no disclosure a low ability examinee probably will have answered incorrectly more of the disclosed items than a more able examinee.

5. A1010.L2. This is another low ability condition, but more extreme than condition 4. In this case an examinee received disclosed information only if the examinee's raw score was less than or equal to 110, which is about 18 points less than the raw score mean of the total group. This condition is intended to reflect a situation where perhaps only failing examinees are given the opportunity to examine the items in form A.

6. A5010. In this condition 50% of the anchor test items were randomly selected for disclosure and 10% of the examinees received the direct benefit of the disclosure. This condition represents an increase in the degree of exposure of the anchor test due entirely to a substantial increase in the number of items directly benefiting 500 form B examinees.

7. A1050. Ten percent of the anchor test items were disclosed and directly benefited 50% of the form B examinees. The increase in exposure in this case is due entirely to an increase in the number of examinees who benefit. Conditions 6 and 7 are specified to compare the results obtained when a large portion of the anchor test benefits only a small portion of the examinee population with the results obtained when a small portion of the anchor test benefits a large portion of the population.

8. A5050. In this condition disclosure of half of the anchor test directly benefited half of the form B examinee population. This condition is intended to represent a relatively high degree of anchor test exposure in the population due to increasing both the proportion of the anchor test disclosed and the proportion of the population that directly benefits from the disclosure.

9. T5050. Under this condition 50% (100 items) of the total test was released to 50% of the population. One-half of the form B examinees received direct benefits of the disclosure. By randomly selecting the 100 disclosed items one would expect approximately 15 to be anchor test items and approximately 85 to be nonanchor test items. It turned out that 17 were anchor test items and 83 were nonanchor test items. This condition could represent a situation where a substantial portion of the form B items were also contained in form A, which was disclosed, and a large portion of the population is affected. This

situation could also occur, however, if there is a major security violation involving form B prior to administration. For example, a copy of form B might be stolen and distributed to a substantial proportion of the examinee population.

10. N5050. Under this condition 50% of the nonanchor test items directly benefited 50% of the examinee population. None of the anchor test items were released, which represents the difference between this condition and the previous one. This situation could occur if many of the items in form B were also contained in form A and only the nonanchor test items in form A were disclosed. The situation could also occur, however, through a violation of security even if all of the nonanchor test items in form B are new items. Consider as an example the loss (possibly by theft) of the new items in form B occurring very late in the development process of the test.

These various disclosure conditions are intended to represent different situations of actual disclosure. No attempt was made to control the difficulty level of the disclosed items or the ability level (except in conditions A1010.L1 and A1010.L2) of the examinees receiving the benefit of the disclosed items. The average difficulty values for the sets of disclosed items and the average raw scores of the subgroups of examinees benefiting from the disclosed items are presented in Table 1. The average difficulty value for all 30 items in the anchor test is .64 and the average raw score for the entire (n=5000) examinee population is 128.7. The average difficulty of the entire 200-item test is also .64.

Insert Table 1 about here.

Results and Discussion

As expected, the equated score means increased, relative to baseline, when anchor test items received some exposure in the form B population. The raw score and equated score means and standard deviations for baseline and all disclosure conditions are presented in

Table 2. When none of the anchor test items were released (condition N5050) the equated score mean decreased slightly compared to baseline.

Insert Table 2 about here.

It is also informative to evaluate the results by examining the percentage of examinees who would pass, based on a specified passing score, and the raw score which converts to the passing score (referred to here as the minimum raw score to pass) under the various disclosure conditions. The minimum raw score to pass is included in Figure 1 with the equated and raw score means, and the percentage passing is included in Figure 2. For purposes of this study the passing score was specified as 120 on the equated score scale. Under the baseline condition, approximately 71% of the examinees would pass at this cut score. Because of the almost identical results obtained for the first five disclosure conditions, these results were consolidated into two "conditions" for the purpose of presentation in Figure 1 and Figure 2; the results for conditions A1010.1, A1010.2 and A1010.3 are represented by A1010, and the results for conditions A1010.L1 and A1010.L2 are represented by A1010.L. The minimum raw score to pass for each of these five conditions is 120. The range for percent passing for these five conditions is only one-half of one percent.

Insert Figure 1 and Figure 2 about here.

In the context of this research it appears that when the anchor test receives only a small degree of exposure in the examinee population the results are only slightly different than when no items are disclosed. The raw score and equated score means and percent passing for conditions A1010.1, A1010.2, and A1010.3 are all only slightly greater than the same statistics under baseline. The similarity of results among these three conditions suggests

that the results obtained under these conditions are representative and not due to sampling fluctuations.

The results obtained under the two low ability conditions are very similar to the results obtained under the first three conditions. It is possible that if the degree of anchor test exposure was greater among low ability examinees, the results would be more pronounced. The results obtained here may be somewhat surprising considering that the level of ability of one of the groups was much lower than the baseline group and the disclosed items were, on average, relatively difficult items, hence the "potential" for greater disclosure effects seemed evident. The degree of anchor test exposure occurring in the two low ability conditions specified in this study is apparently not sufficient to produce substantial increases in the raw and equated score means.

There are at least two factors which, if modified, would probably produce slight variations in the results obtained. For example, the differences in percent passing would most likely change slightly if a passing score other than 120 were specified. Also, the results might be more pronounced if the anchor test constituted a larger proportion of the total test. In the test used for this study, the anchor test was 15% of the total test; 10% of the anchor test was only 1.5% of the total test.

The fact that the results are greater under the A5010 condition than under the A1050 condition might suggest that the proportion of examinees receiving direct benefit of disclosed items has a greater influence on the results than the proportion of anchor test items released. Such a conclusion would probably be premature, however. The differences in results are not great and would likely be sensitive to changes in factors like the specified passing score and the relative size of the anchor test as discussed above.

The positive relationships among increases in raw and equated score means, percentage passing and increased exposure of the anchor test items is reflected in the results of the first eight conditions (all conditions with prefix "A"). The results are greater under A5050

than under either A1050 or A5010; the anchor test received more exposure under A5050. A similar relationship holds between the 10-50 conditions and the 10-10 conditions.

The most extreme results obtained in this research occurred under the two most extreme conditions of disclosure that were studied. Under T5050, when about half of the anchor test items and about half of the nonanchor test items were disclosed, the average raw score increased 18 points, and the average equated score increased almost 12 points. The increase in the passing rate was a substantial 13 percentage points over baseline and, perhaps surprisingly, only 1.5 percentage points higher than condition A5050, where only the anchor test items were released. The minimum raw score to pass increased by two points over baseline.

Under condition N5050, where none of the anchor test items were released, the average raw score jumped by almost 15 points but the equated score mean and the passing rate were very similar to the baseline data. The minimum raw score to pass increased by a relatively large 13 points over baseline. In another study, Lenel and Gilmer (1986) examined the effects on the results of Tucker equating of multiple keying several nonanchor test items. In that study the correct answers for the selected items were placed in all examinees' records. The equated score means for the experimental conditions were found to be no different than the equated score mean from no multiple keying which is similar to the result reported here.

The more complex nature of these results is represented by the unanticipated changes in the minimum raw score to pass across the disclosure conditions. As the exposure of the anchor test items, and only the anchor test items, increases, the minimum raw score to pass declines. When both anchor test and nonanchor test items were disclosed the minimum passing score increased slightly compared to the passing score at baseline. The required score to pass increased substantially over baseline when only nonanchor test items were disclosed.

A possible explanation for these changes in passing score may involve the relationship between examinee performance on the anchor test and performance on nonanchor test items. The correlation between the anchor test scores and scores on the nonanchor test items declines when items from either set, but not both, are disclosed. The correlation between these two sets of items is .65 for baseline, .59 for A5010, .47 for A5050, and .39 for N5050. The correlation increases when both anchor test and nonanchor test items are disclosed; under condition T5050 the correlation was .83.

The test appears to be more difficult than baseline when only anchor test items are disclosed (fewer correct answers are needed to pass) and easier than baseline when nonanchor test items are disclosed (more correct answers are needed to pass). One of the functions of the anchor test is to provide an assessment of the ability of the examinees relative to another group that was administered the same anchor test. The nonanchor test items can then be used to evaluate the overall difficulty level of the test relative to another test containing the anchor test. When only anchor test items are disclosed, examinees appear to be more able than under baseline. This increase in ability, however, does not carry over into the nonanchor test items. The examinees performed the same on the nonanchor test items as under baseline. This improvement in only the anchor test items results in the overall test appearing to be more difficult. When only nonanchor test items are disclosed, the ability level of the group is the same as under baseline but, because of increased performance on the nonanchor test items, the overall test appears to be easier. When both anchor test and nonanchor test items are disclosed the increase in group ability is also reflected in the nonanchor test items and the minimum raw score to pass might be expected to be similar to baseline. In this study the passing score is slightly greater than baseline possibly because of the substantially greater number of nonanchor test items than anchor test items disclosed.

The fairness issue. As indicated earlier in this paper, the proponents of test disclosure argue that disclosure is a matter of simple fairness to the examinees. In the context of

occupational licensing exams, however, a major concern also rests on the issue of fairness to the public, the consumers of the services provided by the licensed applicants. Of course, the test should not be the sole factor on which to base licensing decisions, but the tests "are intended to make a significant contribution toward reliably separating applicants who are competent to provide consumers with safe, effective service from those who are not" (California Department of Consumer Affairs, Central Testing Unit, 1983, p. II-1). The 1985 edition of the Standards for Educational and Psychological Testing states: "The primary purpose of licensure and certification is to protect the public" (American Psychological Association, 1985, p. 63). It appears, then, an important issue requiring attention is that, under test disclosure, the probability of licensing unqualified applicants increases. Based on the hypothetical passing score of 120 (on the equated score scale) specified for this research, under the A5050 condition, where half of the examinees received direct benefit of half of the anchor test, the number of passing applicants would increase by more than 500 over the no disclosure baseline condition; an increase of more than 10%.

It also appears that the argument that test disclosure is fair to the examinees becomes tenuous when examined from a technical perspective. Table 3 contains some possible pass/fail decisions based on different conditions of disclosure for two hypothetical examinees.

Insert Table 3 about here.

In set one it is assumed that under conditions of no disclosure both examinee A and examinee B have raw scores of 117 which is not sufficient to obtain an equated score of 120, the specified passing score. Neither A nor B would pass. This is represented in the raw scores and equated scores across from baseline in set one. When it is assumed that examinee A receives direct benefit from all of the disclosed items and examinee B receives

no benefit of any disclosed items, A obtains an equated score high enough to pass and B does not, except when the anchor test receives substantial exposure in the population as in A5050. The result under condition A5050 is interesting because it suggests that B's equated score can be elevated to passing status *even when* B receives no knowledge of the disclosed anchor test items.

In set three, when half of the full test or half of the nonanchor test items are disclosed, it is assumed that examinee A receives direct benefit of 20 disclosed items and examinee B receives direct benefit of five disclosed items. Both examinees have baseline (no disclosure) raw scores of 117 and would not pass. Under the T5050 and N5050 disclosure conditions A would now pass and B would not.

It also seems possible for the change in passing status to be reversed under conditions of major disclosure. Instead of some examinees passing who would otherwise fail, some may fail under disclosure who would otherwise pass. Based on the assumptions made for set six, both examinees obtain passing scores under baseline and under T5050 only examinee A would pass; under N5050 neither A nor B would pass.

In many licensing programs it is likely that the pass/fail decision is based on other criteria in addition to a major test. A licensing board may decide to specify both conditional and unconditional passing scores. For example, examinees may fail if their equated scores are less than 120; if their equated scores are between 120 and 139 (inclusive), they will pass pending their performance on other criteria; if their scores are 140 or higher, they will pass regardless of their performance on other criteria. The possibility seems to exist for the results of this process to be affected by test disclosure. In Table 3, the results for set two show a situation where both examinees A and B would pass conditionally under no disclosure; under disclosure A would pass unconditionally but B would still be required to perform satisfactorily on the other criteria. Similar results are obtained in set four. In set five examinee B would be required to show satisfactory

performance on other criteria under disclosure which he would not be required to do under no disclosure.

Detectability. Can the effects of test disclosure be detected during the scoring and equating process? If disclosure effects are identifiable and can be isolated, then it may be possible to statistically adjust or control for the effects upon determining final equated scores. The likelihood of this is extremely remote, however. Small changes in many of the important test statistics occur regularly in all testing programs across administrations--even when there is no disclosure. Unless there is major exposure of the disclosed items in the population, it is likely that the effects of disclosure will be small and subtle. In the first year or so following a policy change providing for item disclosure, the disclosed materials will probably receive only slight exposure in the population. The major threat lies, however, in the small but steady increases in the effects of disclosure over time. Such small but steady changes in the score distributions could result in substantial undetected inequities after a few years as the exposure of disclosed items tends to increase.

Alternative equating models. The equating model employed in this study uses an internal anchor test; raw scores on the anchor test are included in the examinees' raw scores. Another possibility is to develop an anchor test that is used in both forms A and B but does not count in the examinees' raw scores. This is an external anchor. Both the California and New York testing laws require disclosure of only those items contributing to raw scores (Brown, 1980; Greer, 1984a). Testing programs that presently use an internal anchor would be required to make significant and far-reaching changes to convert to an external anchor. Although conversion would probably be a major upheaval, it is not impossible. A more immediate problem, however, lies with some test sponsors' strong feelings regarding the possible ethical issues that are raised when examinees are asked to spend time on items that do not contribute to their scores. It is also possible that future laws will require disclosure of all items, even those contained in an external anchor test.

There are also equating procedures based on item response theory (IRT). But IRT equating requires a calibration administration to estimate the item parameters prior to assembling the anchor test. Often, the items to be calibrated are administered with operational items. Persons opposed to equating through an external anchor would have similar concerns about IRT procedures. Additionally, IRT equating may be overly sensitive to violations of the unidimensionality assumption and to differences in the abilities of the groups taking the forms to be equated (Skaggs & Lissitz, 1986).

At this stage in the development of equating methods there does not appear to exist a "best" method in the sense of satisfying all the psychometric requirements, the perceived ethical issues, and the legislative concerns of test disclosure. Until such a model is developed, the alternative is to compromise and use less than "best" procedures to equate scores under requirements of disclosure. The question then becomes--what degree of compromise is acceptable?

Conclusions

The primary purpose of this research was to examine some of the specific effects of test item disclosure on the resulting equated scores and pass rates in the context of policies and legislation related to test disclosure.

The results suggest that the effects of disclosure are dependent on the nature of the released items. For the test and the disclosure conditions examined in this study, it appears that when only anchor test items are disclosed the passing rate increases. The increase occurs not only because the equated score distribution moves up the scale but also because of the unanticipated decrease in the minimum raw score required to obtain a passing equated score.

When only the anchor test items are disclosed the pass/fail decisions on some examinees seem to depend on whether those examinees receive direct benefit of the disclosed items, even when there is only a small degree of exposure of the anchor test items. Examinees

who receive direct benefit of the released items obtain increased raw scores and equated scores. All examinees have higher probability of passing--but the increase is greater for those who receive direct benefit of the disclosed items than for those who receive no benefit of the disclosed items. The increase in the probability of passing is even greater as the anchor test items receive more exposure in the population.

When the new test form contains disclosed items and some of these items make up the anchor test and others are not part of the anchor test, the equated scores tend to move up the scale and more examinees pass. In this study, the minimum raw score required for passing also increased; this could be related, however, to the fact that many more nonanchor test items were disclosed than anchor test items. Even though it is likely that more people will pass under this type of disclosure, it is also possible that some examinees, those who receive no knowledge from the released items, will fail when they would have passed with no disclosure.

When the new test form contains disclosed items, none of which are in the anchor test, the equated score mean and the percent passing remain relatively stable compared to baseline. But overall the test appears to be an easier test and, therefore, the required raw score to pass increases. This implies that some examinees who receive little or no benefit from the disclosed items may not pass when perhaps they would have passed with no disclosure.

Two points deserve to be emphasized regarding the issue of fairness. First, in occupational licensing testing fairness implies protection of the public. If a licensing board's pass/fail criteria have been validly and reasonably determined, then the changes in pass rates which appear to be possible under test disclosure should at least raise the issue regarding adequate protection of the public.

Second, the argument that test disclosure is a matter of simple fairness to the examinee is questionable. The results of this research suggest that the effects of test disclosure are neither simple nor fair.

The psychometric options available for test equating at the present time are limited. All methods appear to be sensitive to disclosure, have ethical drawbacks or suffer from other technical limitations. If a compromise is to be reached, it may well be based on a cost-benefit analysis. Strenio writes:

> If we could quantify the relative costs and benefits of open versus secure tests, the question would still remain of how we choose to weigh factors and what we as a society value enough to pay for. (Strenio, 1979, p. 56)

The question may boil down to the degree of unfairness we're willing to accept. Would we accept openness as long as only 10% (or 5% or 2%) of the examinees are tested unfairly? How many licensed but possibly unqualified practitioners will we tolerate?

Finally, perhaps what is needed is for the major parties involved, the test theorists, the proponents of disclosure, and the examinee populations to work toward not only an understanding of their opponents' position but an understanding of the possible ambiguities and implications of their own positions. Possibly then the parties will progress from arguing at each other from different levels to arguing with each other on common ground. Hopefully, this would be only an intermediate step toward achieving more understanding and cooperation.

Table 1

Average Difficulty Level for Disclosed Items and
Average Raw Score for Examinee Groups Receiving Disclosed Items

Condition	No. Items Disclosed	Average Difficulty	No. Examinees Benefiting from Disclosed Items	Average Raw Score*
1. A1010.1	3	.52	500	129.3
2. A1010.2	3	.79	500	128.7
3. A1010.3	3	**	500	129.3
4. A1010.L1	3	.52	500	114.3
5. A1010.L2	3	.52	500	100.3
6. A5010	15	.72	500	129.3
7. A1050	3	.52	2500	128.9
8. A5050	15	.72	2500	128.9
9. T5050	100	.64	2500	128.9
10. N5050	85	.65	2500	128.9

Note: The average difficulty of the anchor test is .64. The average raw score for the 5000-member population is 128.7.

*Average raw score was determined prior to disclosure.
**All 30 anchor test items were disclosed in subsets of three items to 10 groups of 50 examinees each.

Table 2

Raw Score and Equated Score Means and Standard Deviations for Baseline and Disclosure Conditions After Item Disclosure

Condition	Raw Score Mean	Raw Score SD	Equated Score Mean	Equated Score SD
BASELINE	128.7	16.3	128.5	17.2
1. A1010.1	128.8	16.3	129.0	17.2
2. A1010.2	128.7	16.3	128.7	17.2
3. A1010.3	128.8	16.3	128.8	17.2
4. A1010.L1	128.8	16.2	129.0	16.9
5. A1010.L2	128.9	16.0	129.1	16.6
6. A5010	129.1	16.3	129.9	17.6
7. A1050	129.4	16.3	131.2	17.2
8. A5050	130.8	15.9	135.6	18.2
9. T5050	146.4	22.3	140.0	18.9
10. N5050	143.2	20.0	127.7	17.4

Figure 1

Equated and Raw Score Means and Minimum Raw Score to Pass for Several Disclosure Conditions

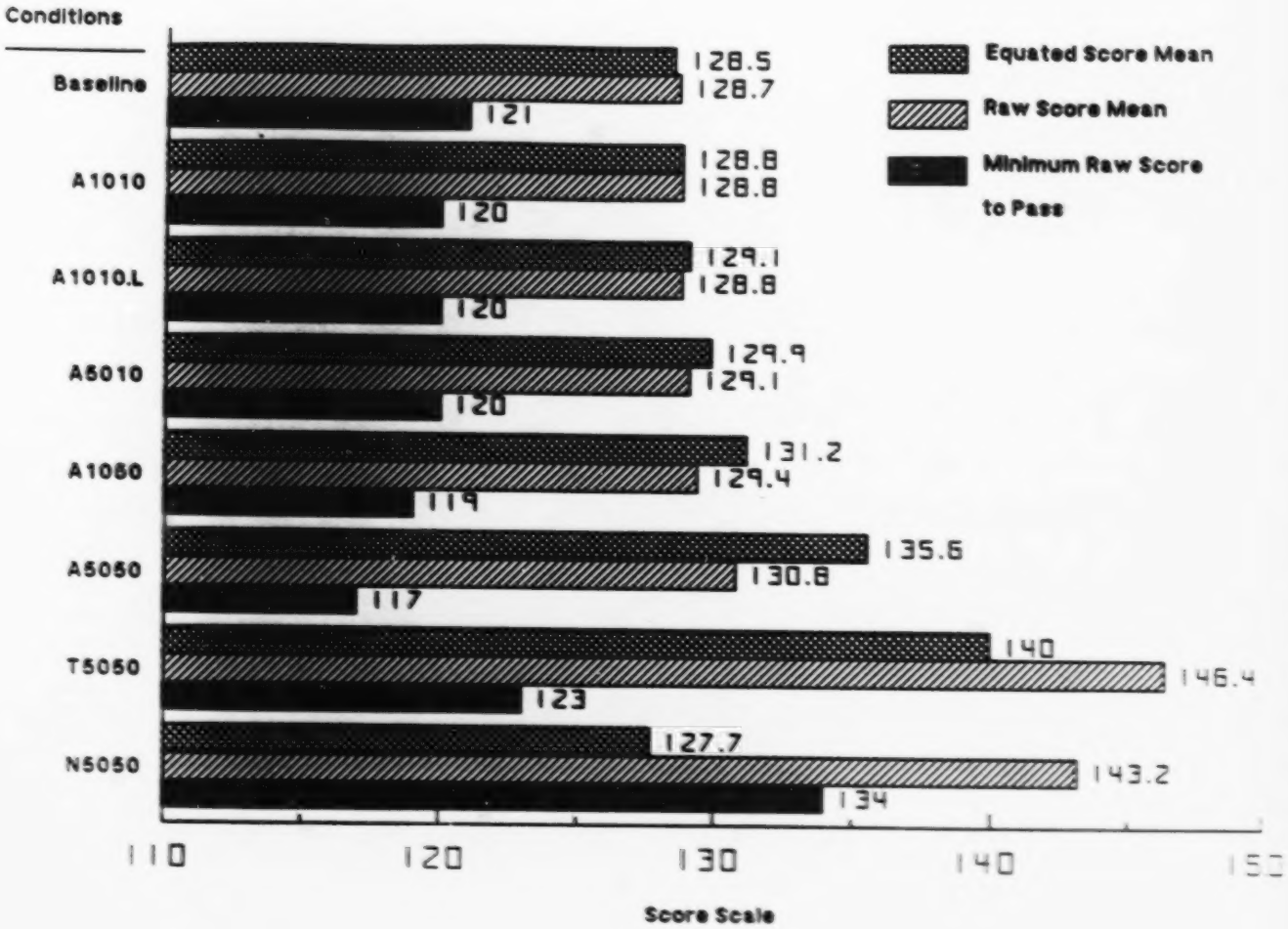

Figure 2

Percent Passing and Minimum Raw Score to Pass for Several Disclosure Conditions

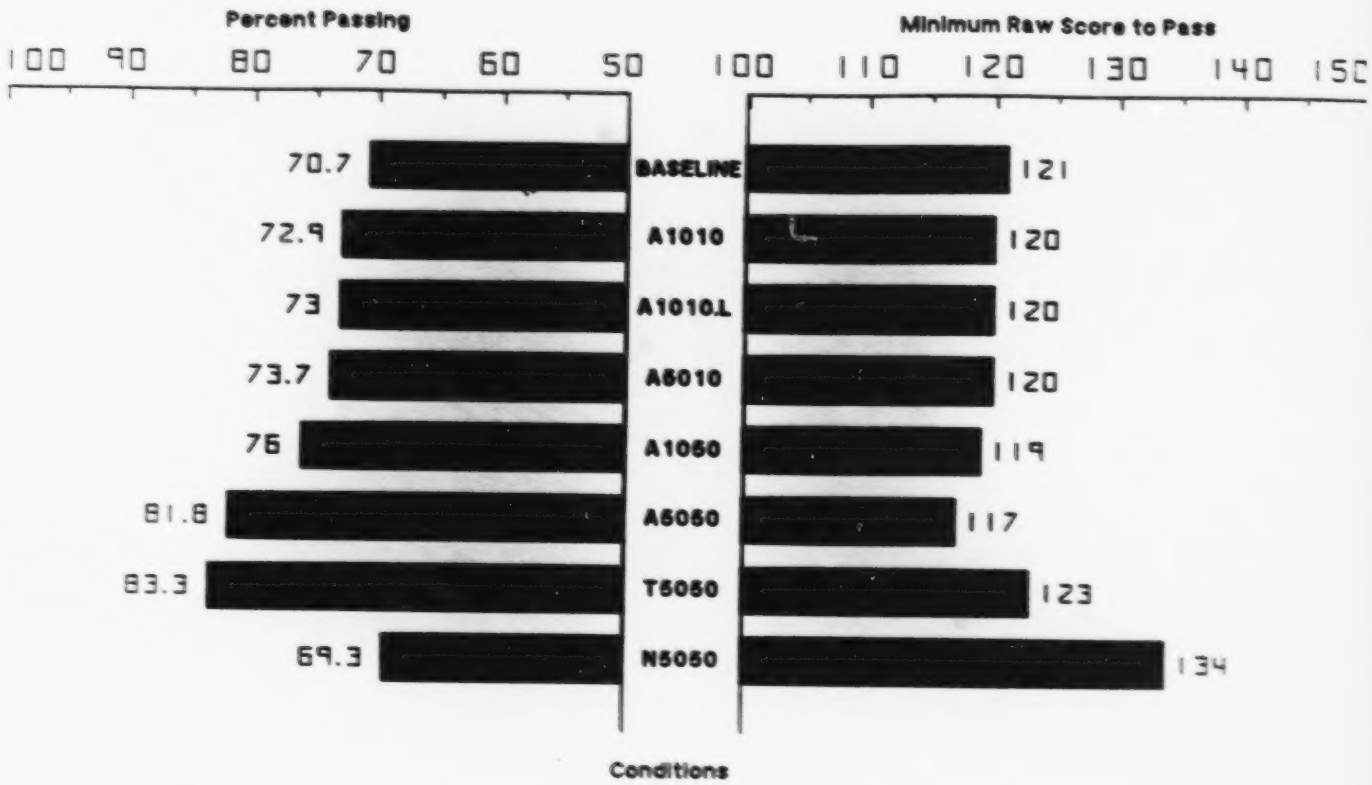

Table 3

Raw Scores, Equated Scores and Pass/Fail Decisions Across Several Disclosure Conditions and Three Different Assumptions of the Direct Benefits of Disclosed Items

Assume:		Condition	Raw Score Examinee A	Raw Score Examinee B	Equated Score Examinee A	Equated Score Examinee B
Examinee A receives direct benefit of all disclosed items. Examinee B receives no benefit of disclosed items.	Set 1	BASELINE	117	117	116	116
		A1010	120	117	*120*	116
		A1010L	120	117	*120*	117
		A5010	132	117	*133*	117
		A1050	120	117	*121*	118
		A5050	132	117	*137*	*120*
	Set 2	BASELINE	125	125	*125*	*125*
		A5010	140	125	*142**	*125*
Examinee A receives direct benefit of 20 disclosed items. Examinee B receives direct benefit of 5 disclosed items.	Set 3	BASELINE	117	117	116	116
		T5050	137	122	*132*	119
		N5050	137	122	*122*	109
	Set 4	BASELINE	137	137	*137*	*137*
		N5050	157	142	*140**	*127*
	Set 5	BASELINE	140	140	*140**	*140**
		N5050	160	145	*142**	*129*
Examinee A receives direct benefit of 9 disclosed items. Examinee B receives direct benefit of 1 disclosed item.	Set 6	BASELINE	121	121	*120*	*120*
		T5050	130	122	*126*	119
		N5050	130	122	116	109

Note: Underlined italicized numbers represent passing scores where the criterion passing score is set at 120. The asterisked (*) scores represent unconditional passing scores where the criterion unconditional passing score is set at 140.

References

American Psychological Association. (1985) Standards for educational and psychological testing. Washington, DC, author.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.), 508-600. Washington, DC: American Council on Education.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland & D. B. Rubin (Eds.), Test equating, 9-49, New York: Academic Press.

Brown, R. (1980). Searching for the truth about "truth-in-testing" legislation, a background report. (Report No. 132). Denver, CO: Education Commission of the States.

Brennan, R. L., & Kolen, M. J. (1986, April). Practical issues in linear equating using the common item nonequivalent populations design. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

California Department of Consumer Affairs, Central Testing Unit. (1983). What a licensing board member needs to know about testing. Sacramento, CA, author.

Greer, D. G. (1984a). "Truth-in-testing legislation," an analysis of political and legal consequences, and prospects. (Monograph No. 83-6). Houston: Institute for Higher Education Law and Governance.

Greer, D. G. (1984b). Legal issues in truth-in-testing legislation. The Review of Higher Education, 7, 321-356.

Hale, G. A., Angelis, P. J., & Thibodeau, L. A. (1981, April). Effects of item disclosure on TOEFL performance. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Kolen, M. J. (1985, April). Comparison of methods for linear equating under the common item nonequivalent populations design. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Lenel, J. C., & Gilmer, J. S. (1986, April). The effect of keying all options correct on equating functions and scores. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: relevant issues and a review of recent research. Review of Educational Research, 56, 495-529.

Strenio, A. (1979). The debate over open versus secure testing: a critical review. (Staff circular No. 6). Cambridge, MA: The Huron Institute, National Consortium on Testing Project.